Product Features Extraction and Categorization in Chinese Reviews

Shu Zhang, Wenjie Jia, Yingju Xia, Yao Meng, Hao Yu
Information Technology Laboratory
Fujitsu Research & Development Center
Beijing, China
e-mail: {zhangshu, wj_jia, yjxia, mengyao, yu}@cn.fujitsu.com

***Abstract*—With the growing interest in opinion mining from web data, more works are focused on mining in English and Chinese reviews. Product features extraction and categorization are very important for feature level opinion mining. In this paper, we propose a supervised product features extraction method, regard it as an entity recognizing process, and hope to transfer the effective NER techniques to solve this problem. We propose an unsupervised method to group the product features, mine the association of the product features from the intra and inter relationship. With experiments on Chinese reviews, the results show that proposed techniques for product features extraction and categorization are proved effective and promising. The opinion words are very important features both in features extraction and categorization.**

Keywords- opinion mining; product features; categorization

I. INTRODUCTION

With the growing of Web 2.0 platforms such as blogs, forums and various other types of social media, it becomes possible for people to find useful experience and advice from reviews or comments on products or services. Opinion mining has been proposed to analyze reviews and extricate people from wading through a large number of opinions to find their interest. People usually pay more attention to some aspects of product, it is useful to extract and analyze product features from the reviews. Product features extraction belongs to feature level opinion mining, which is finer-grained opinion mining compared with document and sentence level opinion mining.

In recent years, some feature level opinion mining systems have been presented to capture reviews' opinions on different product aspects. Opinion Observer [1] focuses on online customer reviews and provides the visual comparison of customer opinions of products on various product features. Red Opal [2] offers to find products based on features and scores each product on each feature. This information is useful to both potential customers and product manufacturers.

In reviews, people usually describe the same product features by different words. It is necessary to group them together in order to analyze the overall sentiments on one product feature. For example, "photo", "picture" and "image" all refer to the same aspect in digital camera reviews and should be grouped together, otherwise it is too detailed and tedious for customers and merchants to read and summarize all these product features. It is also infeasible and time-consuming to group the massive product features manually.

This paper focuses on extraction and categorization of product features in Chinese reviews. We propose a supervised product features extraction method, regard it as an entity recognizing process, and hope to transfer the effective NER techniques to solve this problem. We propose an unsupervised method to group the product features, mine the association of the product features from the intra and inter relationship. In this stage, we focus on finding the good indicators to reveal the association of product features and show how their influence on the performance of grouping results.

The remainder of the paper is organized as follows: Section 2 describes the related work on extraction and categorization of product features. Section 3 describes product features extraction. Section 4 presents product features categorization. Section 5 gives the experiments and results. Finally, Section 6 summarizes this paper.

II. RELATED WORK

The techniques for identifying product features are primarily based on unsupervised mining. The most representative research is that of [3]. They adopt association rule mining for extracting nouns as frequent features. Compactness pruning and redundancy pruning are used to filter the incorrect features. Popescu and Etzioni [4] utilize relation-specific extraction patterns with web PMI assessor to assess feature candidates. However, using frequency measure tends to prefer to high frequency features. This leads to the low frequency ones might be missed.

Different from their works, we adopt supervised method to extract product features. We combine frequency, syntax tokens and domain knowledge to find the product features. The importing of domain knowledge is aimed to improve the quality of extraction. With the manually tagged training corpus, we transfer the task of product features extraction into traditional information extraction task using CRFs model.

Grouping product features with similar meaning together is a recent focus in feature level opinion mining. Liu [1] employs WordNet to find synonym groups/sets exist among the features. The coverage of the lexicon is bottleneck of the lexicon-based method. Su [5] proposes a mutual reinforcement approach to clusters product features and opinion words simultaneously and iteratively by fusing both their content information and sentiment link information.

ISBN: 978-1-61208-139-7

The inter link between product features and opinion words are mined to reinforce the clustering quality. Guo [6] constructs latent semantic association model to group words into a set of concepts according to their virtual context documents, then categorizes product features according to their latent semantic structures and context snippets in the reviews. Su [5] and Guo [6] all choose words as the basic smallest units.

Different from their methods, we adopt classical K-Means algorithm to group product features, pay more attention to mine the association of product features. We present morphemes as the smallest linguistic meaningful unit, measure the intra relationship of the product features. We mine different context information to measure the inter relationship of product features.

III. PRODUCT FEATURES EXTRACTION

In feature level opinion mining, the task is to extract product feature associated with its sentiment orientation. The task is typically divided into three main subtasks: (i) identifying product features, (ii) identifying opinions regarding the product features, and (iii) determining the sentiment orientation of the opinions. This paper mainly focuses on the first step to extract product features in Chinese customer reviews.

A. CRFs Model

The product features are mostly noun or noun phrases. In reviews, opinion words mostly appear around the product features in the sentence. The product features are context related, and for a given domain it has the lexical or syntactic similarity. For example:

"相机屏幕大，画面清晰。" (The camera has a big screen, and photo is very clear.)

Here, "屏幕"(screen) and "画面"(photo) are product features. "大"(big) and "清晰"(clear) are opinion words associated with product features.

We transfer the product features extraction to a sequence tagging problem, and hope to utilize effective NER techniques to solve this problem. Another reason for us to adopt the supervised method to implement this task is that the unsupervised frequency-based methods are dependent on the statistic of the corpus, they couldn't execute effectively when given a single sentence.

Conditional Random Fields Model is proposed by Lafferty [7], which has been proved well performance in information extraction field. It has the advantages of relaxing strong independence assumptions made in HMM [8], and avoiding the label bias problem existed in MEMM [9]. We adopt CRFs model to extract product features.

B. Feature Selection

Feature selection has been an active research pattern recognition, statistics and data mining communities. It is often the case that finding the correct subset of features is an important problem. It may significantly improve the performance of supervised learning algorithm.

In this paper, feature selection is based on some criterions: product features are mostly noun or noun phrases, and more appear in an opinion expression. That means the structure and opinion related semantic information are important. So we utilize some shallow semantic features and domain knowledge. The features are shown in the following, which include word, POS and semantic information:

Word information: We consider the neighboring words in a region with the max window 4 in order to get the context information.

POS information: POS is annotated to capture the word-building and simple syntax information. The noun phrase could be exhibited in a neighboring window with the part-of-speech tags--noun, verb, adverb, punctuation, etc.

Semantic information: we utilize some language resource to get the semantic information, such as domain feature lexicon, opinion lexicon, factor words lexicon.

We search whether the word is in product feature lexicon or not, even whether it is the part of an item or not. Because the product features might have the same or similar component in a given domain. For example, "光学变焦"(optical zoom), "数码变焦"(digital zoom), and "变焦镜头"(zoom lens) are all the product features of digital camera domain, they have the same word "变焦"(zoom) as their component.

The appearance of an opinion word or emotional adverb is more likely to indicate the presence of an opinion. As observed, people often like to express their opinions around the product feature. In product reviews, especially in Chinese reviews, people like to express their opinion in short and simple sentence, like the form of "product feature" + "opinion word". The importing of opinion and adverb lexicons aims to utilize more domain knowledge and opinion information. Since we could catch the simple collocation and pattern between the opinion word and the product features in a window by these information.

IV. PRODUCT FEATURES CATEGORIZATION

Product features categorization aims to group product features with similar meaning together. The challenge in product features categorization is how to capture the association among product features from the review. In this paper, we adopt traditional K-Means algorithm to cluster product features, and focus on mining the association of product features. We consider the association from two sides: intra relationship and inter relationship among product features. Intra relationship means the inner linguistic meaningful unit relationship between two product features. The inter relationship means the relationship of context information of two product features.

A. Morpheme Based Intra Relationship

Most researchers choose words as the basic smallest units in opinion mining. With words as basic units, it can't capture the similarity among some product features. For example, we want to measure the intra relationship among product features "电池" (battery), "电源" (power), and "电池续航能力" (battery endurance). Among them, the pair ("电池" (battery), "电池续航能力" (battery endurance)) has an intra relationship for they have the same word "电池" (battery).

The pairs (“电池” (battery), “电源” (power)) and (“电源” (power), “电池续航能力” (battery endurance)) have no intra relationship as they have no same word. In fact, these product features have the similar meaning in reviews.

Looking smaller units than words level, the above three product features all contain the character “电”. This is a good indicator to reflect the association among them. Yuen [10] infers semantic orientation of Chinese words from their association with strongly-polarized Chinese morphemes. The conclusion is that morphemes in Chinese, as in any language, constitute a distinct sub-lexical unit, and have greater linguistic significance than words.

So we choose the morphemes to be smallest linguistic meaningful unit to mine the intra relationship among product features, and calculate the inner characters similarity of product features.

In Chinese, morphemes are mostly monosyllabic and single characters, although there are some exceptional poly-syllabic morphemes like “葡萄”(grape), “咖啡”(coffee), which are mostly loanwords.

In reviews, morphemes reflect the core meaning of product features clearly. For example, “镜”(lens), “屏”(screen) and “像”(photo) are the important component of product features in digital camera reviews.

B. *Opinion Words Based Inter Relationship*

Intra relationship only mines the association among product features from their inner characters components. This information is limited. It is not enough to capture the underlying semantic association of various product features.

In feature level opinion mining, product features and opinion words are basic element. The opinion words mostly appear around the product features in the review sentences. They are highly dependent on each other. It is obvious that surrounding opinion words may play an important role in clustering product features. So we mine the inter relationship among product features utilizing the context information, especially the opinion words associated with product features.

There are hidden sentiment association existing between product features and opinion words. For example, “外型”(shape) and “样子” (appearance), they are not similar on morphemes level, and could not be linked with intra relationship though they refer to the same aspect in reviews. However, they may be evaluated by similar opinion word “美丽” (beautiful). The opinion words describing this aspect of “appearance” are often using the words “美丽” (beautiful), “时尚”(fashion), “流行的”(popular) etc. So the opinion words around the product features really contain the semantic information to reflect the inter relationship among product features.

C. *Representation*

Product features categorization is conducted by representing each data object instance by a feature vector. We represent an instance as a set of following features.

Morphemes units M: all the characters contained by x_i.

Opinion words units O: only the opinion words in the given window size {-*t*, *t*} are considered.

The weight of each features units f_j^i is calculated by Mutual Information.

$$PMI(f_j^i, x_i) = \log_2 \frac{P(f_j^i, x_i)}{P(f_j^i)P(x_i)} \qquad (1)$$

Where, $P(f_j^i, x_i)$ is the joint probability of x_i and f_j^i co-occurred in the corpus. $P(f_j^i)$ is the probability of f_j^i occurred in the corpus. $P(x_i)$ is the probability of x_i occurred in the corpus. The ratio is a measure of the degree of statistical dependence between the x_i and f_j^i.

V. EXPERIMENTS

In this section, we evaluate the proposed methods and analyze the performance of product features extraction and categorization in detail.

A. *Performance of Product Features Extraction*

This experiment is conducted on the corpus provided by the COAE (The first Chinese Opinion Analysis and Evaluation), which was held in 2008, aims to enable researchers to participate in large-scale experiments and evaluations, make each researcher’s result comparable and promote the related technique in Chinese opinion analysis.

The corpus contains automobile and electronic domains, with about 1,500 sentences each. All product features have been annotated by human.

The precision, recall and F-measure will be used to measure the performance. We adopt strict matching, which means the results submitted by systems are exactly same with the human labels.

Table I and Table II present the evaluation results. For the comparison with others, we also give the Avg. and Max. values in the task. There are 13 participants in this task. Our system is named as FRDC. We aim to testify the performance of proposed method and its capability of domain transplant.

TABLE I. RESULTS OF PRODUCT FEATURES EXTRACTION

RunID	**Precision**	**Recall**	**F-measure**
FRDC	0.3798	**0.4172**	**0.3976**
Avg	0.2877	0.2270	0.2331
Max	0.5641	0.4172	0.3976

TABLE II. DETAIL RESULTS ON DIFFERENT DOMAINS

RunID	**Precision**	**Recall**	**F-measure**
Automobile	0.2435	0.3326	0.2811
Camera	0.3512	0.3563	0.3537
Phone	0.3920	0.3539	0.3720
NoteBook	0.3782	0.3880	0.3830

In Table I, compared with the average and maximum value gotten in the COAE, the value of FRDC in F-measure proves that CRFs-based feature extraction is feasible and valid. Our system's precision and recall are similar and not inclined to one parameter excessively, which means our method is more practical and feasible.

Table II shows the detail results of performance on different domain. The test data include automobile and electronic domains. The electronic domain has the Camera, Phone and Notebook sub-domain. The performance on electronic domain is better than that on automobile domain on all parameters. However for the sub-domain on electronic domain the performance is similar. So it could be concluded that the performance of the system is affected by the domain, but is not sensitive. When the difference between the two given domain is not significant, the performance is similar. One reason for low performance on automobile domain might be caused by that the product features is longer and more complex.

B. Performance of Product Features Categorization

With the limited of human efforts and time, we testify the performance of product features categorization on digital camera domain. The corpus contains about 22,000 posts, extracted from the review websites. Three humans categorize product features into the categories, and we choose the label agreed by at least two humans as the standard. The detail of the corpus is shown in Table 3. We suppose that product features are extracted correctly.

TABLE III. CATEGORIZATION EVALUATION SET

Category	**Number of product features**
Lens	56
Screen	62
Appearance	110
Battery	18
Photography	76
Total	322

The performance of product feature categorization is evaluated using the measure of Rand Index. It is a measure of cluster similarity.

$$Rand(P_1, P_2) = \frac{2(a+b)}{n \times (n-1)} \tag{2}$$

Where, P_1 and P_2 respectively represent the partition of an algorithm and manual labeling. The agreement of P_1 and P_2 is checked on their $n \times (n-1)/2$ pairs of instances, where n is the size of data set *D*. For each two instance in *D*, P_1 and P_2 either assigns them to the same cluster or to different groups. Let *a* be the frequency where pairs belong to the same group of both partitions. Let b be the frequency where pairs belong to the different group of both partitions. Then Rand Index is calculated by the proportions of total agreement.

In our experiment, *D* contains the product features words in the pre-constructed evaluation set. Partition agreements between the pairs of any two product features are checked automatically. This measure varies from 0 to 1. The score of 1 is the best.

We first testify the performance of the proposed techniques from two perspectives:

1 The effectiveness of inducing morpheme as features to measure the intra relationship among product features.

2 The effectiveness of opinion words as feature to measure the inter relationship among product features.

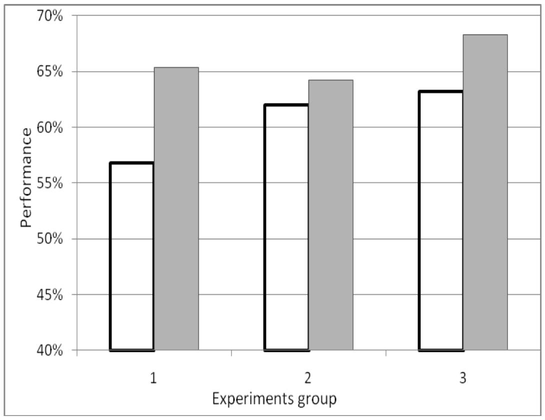

Figure 1. Different feature chosen result.

In Figure 1, the experiment of No.1 group compares the performance of method using full context as features with that of opinion words. It is only considered the inter relationship among product features, no consideration of intra relationship. The left column is the method using full context as features, which is much less than that of opinion words as features (the right column) in accuracy value. That proves the opinion words are good at indicating the semantic similarity of product features associated with them. Compared with opinion words, the full context more likely induce some noise information.

Based on No.1 group, we induce intra relationship measurement. The experiment of No.2 group induces the intra relationship measurement based on word level. The experiment of No.3 group measures the intra relationship with morphemes. No.3 group achieve better accuracy than both No.2 group and No.1. That proves morpheme features are more effective than word features. The inducing of morpheme features to measure intra relationship enhances the performance.

VI. CONCLUSION

In this paper, we probe into the problem of product features extraction and categorization. We propose CRFs-based method to extract product features in reviews. We propose an unsupervised product features categorization method. With the experiments in Chinese reviews, the proposed methods achieve better performance. CRFs-based product features extraction is effective and feasible. Morphemes and opinion words are proved to be the important features to capture the semantic similarity among product features in process of product features categorization.

However, the methods are only tested on Chinese customer reviews. We will conduct experiments on different languages and domains in future work.

REFERENCES

[1] B. Liu, M.Q. Hu, and J.S. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proc. International World Wide Web Conference, pp. 342–351, 2005.

[2] C. Scaffidi, K. Bierhoff et al. "Red Opal: Product-Feature Scoring from Reviews," Proc. 8th ACM Conference on Electronic Commerce, pp.182–191, 2007.

[3] M.Q. Hu, and B. Liu. "Mining Opinion Features in Customer Reviews," Proc. American Association for Artificial Intelligence, pp. 775–760, 2004.

[4] A. Popescu, and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. on Empirical Methods in Natural Language Processing, pp. 339–346, 2005.

[5] Q. Su, X.Y. Xu et al, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th International Conference on World Wide Web, pp. 959–968. 2008.

[6] H.L. Guo, H.J Zhu et al, "Product Feature Categorization with Multilevel Latent Semantic Association," Proc. International Conference on Information and Knowledge Management, pp.1087–1096, 2009.

[7] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. International Conference on Machine Learning, pp. 282–289. 2001.

[8] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77 (2), pp. 257–286, 1989.

[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," Proc. on International Conference of Machine Learning, pp. 591–598, 2004.

[10] W.M Yuen Raymond et al, "Morpheme-based Derivation of Bipolar Semnatic Orientation of Chinese Words," Proc. 20th International Conference on Computational Linguistics, pp. 417–424. 2004.